Virtual Fitting Room with Spoken Dialogue Interface

Tatsuya Kawahara *Katsuaki Tanaka* *Shuji Doshita*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

ABSTRACT

We address an effective application of spoken dialogue interface by combining with virtual space technologies. Virtual space enables us to do what cannot be done in the real world with feeling as if doing in the real world. Spoken dialogue systems are to realize our intention by understanding expressions that include concept and a nuance. We clarified advantages and disadvantages of speech interface in a multi-modal virtual space, then designed a virtual fitting room, where the speech interface works effectively and the dialogue is enhanced by multi-modal interaction. The system supposes that a user is in a fitting room. The user can select his favorite clothes interactively by stating his preference and checking the virtual mirror. It was observed that the combination of spoken language input and virtually-real image output realizes a natural and robust interaction, thus giving better user satisfaction than conventional platforms.

1. INTRODUCTION

While numerous multi-modal interfaces are being designed and implemented, speech input is not necessarily used effectively as expected because it may be less competitive than other modalities such as pointing devices with respect to the accuracy and efficiency. When multi-modal systems assume tasks of input or manipulation of information using a computer, speech input can be well complemented by other modalities[1]. In most of virtual reality (VR) systems that are also multi-modal systems, however, pointing devices such as special gloves and sticks are commonly used to specify the users' intention.

Therefore, we first discuss which aspects of speech interface are vitally useful or more advantageous than those of other input devices. Then, we present a new application of spoken dialogue interface in a virtual space environment[2]. We have designed and implemented a virtual fitting room, where spoken dialogue processing is integrated with image processing and the speech input is effectively used to specify users' preferences for their clothes.

2. SPOKEN DIALOGUE IN VIRTUAL SPACE ENVIRONMENT

2.1 Virtual Space Environment

In this paper, virtual space is defined as an environment where a human can interact in the same manner as in the real world to perform what cannot be done in the real world. For example, "virtual city" enables users to walk and see computer-synthesized streets and shops. "Virtual traveling environment" makes users feel as if they are traveling to distant countries without actually going out.

As a vital factor of natural and preferable interfaces, the manner of interaction must be the same as that in the real world, and so as the (visual and audio) feedback information. Moreover, the virtual space environment realizes operations that are free from the constraints of time and space in the real world. For example, it can provide infinite stock, move objects to distant places in a second, and even play back the (virtual) time to retry.

2.2 Effect of Speech in Virtual Space

In most of the conventional virtual reality (VR) systems, common modalities for interaction are pointing devices such as data gloves and visual feedback such as a CG screen[3]. Only a few systems adopt speech input to specify objects such as "put that there"[4]. But such commands can be easily replaced by gestures or pointing devices.

Thus, it is vital to find an effective use of the speech input, specifically to clarify functions that other modalities can not or hardly provide. These are classified as follows.

1. To specify invisible objects
   Since speech commands can specify objects that are not visible on a screen or in a virtual space, it is possible to set up virtual chests or storehouses of enormous stock and allow access to items there.
2. To specify attributes
   Speech commands make it easy to specify colors and sizes, or even some nuances, for example, "a red one" or "a bit brighter".
3. To command in hands-free environment

Speech input does not require users to carry any devices, nor constrain their hand movement if distant microphones are used. This feature enhances the virtual reality and naturalness of the user interface.

4. To select from large choices
   Users can specify objects directly by their names via speech without searching for the desired one from a menu, whereas it is easy to find and click if the number of choices is less than 10.

2.3 Effect of Multi-Modality to Dialogue

When we incorporate speech recognition to multi-modal interfaces, we must take into account its disadvantages compared with other modalities.

First, speech is less suitable for expressing topology and measurements. Other modalities like gestures are essential for such purposes.

Secondly, recognition errors are inevitable. In most cases, the errors lead to unexpected operations and results.

Thirdly, users do not necessarily know the vocabulary and grammar that are acceptable to the system. Thus, they utter out-of-vocabulary (OOV) words or out-of-grammar (OOG) expressions that can hardly be dealt with by systems and lead to errors.

Although more robust recognition strategies such as key-phrase detection must be explored, we must realize these kinds of errors are inevitable and set up a framework for recovery.

Spoken dialogue systems try to resolve these problems by making confirmations, but frequent confirmations are often troublesome to users. On the other hand, multi-modal response can provide more effective feed-back. If users' utterances are promptly interpreted and reflected in the virtual space such as change of objects or colors, the users can easily understand whether their utterances are appropriately recognized, and correct them if necessary without explicit confirmation by spoken dialogue.

This kind of interaction is also effective to assist users' consideration. Users do not always have a definite preference before dialogue. Multi-modal interaction enables them to try and compare several possible choices, finally leading to satisfactory selection.

3. VIRTUAL FITTING ROOM

Based on the above analysis, we have designed and developed a virtual fitting room by integrating our spoken dialogue interface[5] with image processing under a virtual space platform.

Our particular focus is to realize an interface where (1) users can express their preferences of attributes with spoken language, (2) the system does not impose any constraint on users' movement nor hands.

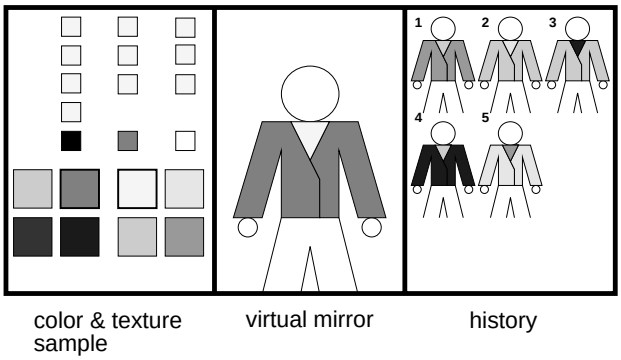

Figure 1. Displays of virtual fitting room

3.1 System Description

Even though on-line shopping is becoming popular, people still want to try on clothes before actually purchasing to make sure the garment fits perfectly. Our virtual fitting room resolves such demands by setting up a virtual mirror that super-imposes the selected clothes into the user's video image of the actual size and motions. The virtual fitting room can prepare an enormous stock of colors and sizes, and the users can instantly change clothes according to their preferences. With spoken dialogue interface, it is possible to specify preferences with spoken language as if speaking to a clerk (though invisible) at a shop.

An outlook of the system is shown in Figure 1.

A user states his preference regarding colors and textures of a jacket or a shirt, such as "I want a red striped jacket." or "I want that a bit brighter." The system recognizes and interprets such speech inputs as color values and texture patterns of the clothes. Then, it selects candidate patterns from a virtual stock that satisfy the condition. Typical colors and possible candidate patterns of textures are presented on a display (left screen in the Figure).

The system also captures users' image with a video camera, and extracts the region of the jacket and shirt. The system then super-imposes the selected color and texture pattern onto the recognized region. This mechanism makes the user feel as if he is trying on the garment in front of a mirror (center screen in the Figure).

As the image is processed in almost real time and projected in real size, the user can move his hands freely as if he is in a fitting room. He will continue the dialogue until a satisfactory one is obtained by changing or correcting previous selections. During the session, several candidate choices that the user actually liked can be stored in the history screen for reference. This makes it possible to compare the current candidate with previous selections. (right screen in the Figure).

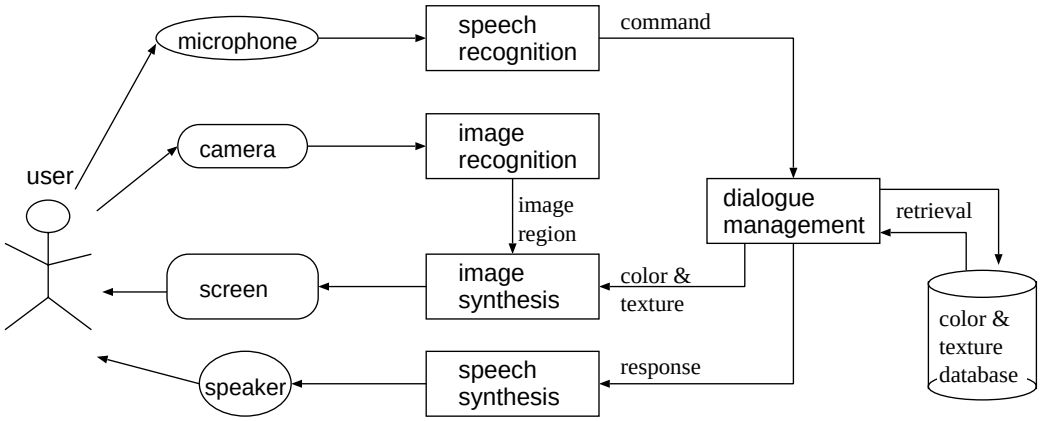

Figure 2. System configuration of virtual fitting room

3.2 User Command Specification

User commands are classified as follows.

- Specification of colors and textures
  - names of colors (e.g.) red, white
  - objective adjectives on colors (e.g) brighter
  - names of texture patterns (e.g.) striped
- Selection from menu screens
  - selecting from the sample screen
  - access to the history screen
- Control
  - undo
  - quit

There are wide variations in the names of colors, such as 'lemon' and 'chocolate', which come mainly from objects with such colors, but it is hard to list all possibilities. So the names of colors are limited to the standard ones only. People may also want to use more subjective adjectives, like 'gorgeous' or 'fresh'. However, the usage and concept of these adjectives are rather individual, thus it is difficult to map them into actual image values. Such words are not handled in this study.

These commands are actually in Japanese language, and the vocabulary size for this task is 83.

3.3 System Configuration

The system architecture and process flow are depicted in Figure 2.

Speech is input from a microphone that is set up at the mirror screen. A microphone array system is desirable to realize full hands-free recognition, but in this application users are expected to stand close to the mirror. Japanese speech recognition modules were developed as a joint project [6]. In this application, we adopted a finite state automaton (FSA) parser. Key-phrases are extracted and converted to a command. A command consists of an object (jacket or shirt), its attributes (e.g. hue of colors) and their values.

A color is defined by three values of hue, saturation and brightness. These are adjusted according to the user specification. So any (full 16M) color output is possible. A texture pattern consists of one main color and sub-colors, whose combination is also affected by the user's preference (simple or showy).

A dialogue manager performs ellipsis analysis using a session history and database query to retrieve candidate color and texture patterns. It also controls the output of image and speech.

Image is input from a camera that is also set up at the mirror screen. Chroma-key synthesis was applied in this study. Users are requested to wear a jacket and shirt of specific mono-tone colors (red and blue). These regions are automatically extracted by the chroma-keys. Original values of brightness are used in the super-imposed image to reflect the shade information of the input image.

4. IMPLEMENTATION AND EVALUATION

The system was implemented with three sets of an 80-inch screen and a projector. Speech and image processing is done at standard workstations. Speech recognition is performed in real time, but there is a few second delay to synthesize and project the image. The video image is processed at about ten frames per second.

An example of the synthesized image is shown in Figure 3, though it is actually video image of real size.

We had 10 people try the system, namely to fit a jacket and shirt freely. There were no serious recognition errors or dialogue crashes observed. The proposed multi-modal system was evaluated by comparing with conventional systems. We asked the subjects to use both systems and to complete questionnaires with respect to the ease of operation and the degree of satisfaction in 5 relative grades.

Figure 3. A shot of virtual mirror

First, to evaluate the effect of speech interface, we compared the system with a menu-and-mouse-based system, where users select with a pointing device from hundreds of color and texture menus. According to the subjective assessment shown in Figure 4, we get conclusions that (1) the menu-and-mouse system is easier for making selections, but (2) the spoken dialogue interface is more satisfying. In the menu selection system, the displayed and selected pattern is projected as it is. This feature makes the operation easy, but some users feel passive when the operation only involves selection from a menu screen. With spoken dialogue interface, however, users can express conceptual specification that includes nuances. So the output image is not predictable and sometimes unexpected. In addition, users are not constrained to search and click menus. Thus, they can focus on the mirror screen and take the initiative in the session. These features are considered to improve the degree of satisfaction.

Next, the effect of the real-size video image is verified by comparing with a PC-screen-based system that makes use of a still image only. As shown in Figure 5, the majority of the subjects answered that the virtual space gave them more satisfaction. The still image projected on a PC screen looks like a small picture, and is not suitable for fitting a new garment. In addition, users simply have to wait during the image processing, while our system constantly updates the current mirror image despite some delay. Keeping the interaction is also significant.

5. DISCUSSIONS

Unlike a drawing system or route-guiding system, the goal of the task in this study is not well defined. It is possible that users will find a much better choice after a very long session. Thus, it is not appropriate to evaluate the system with the recognition accuracy nor the elapsed time. But it is not easy to measure how satisfactory the found clothes are[7].

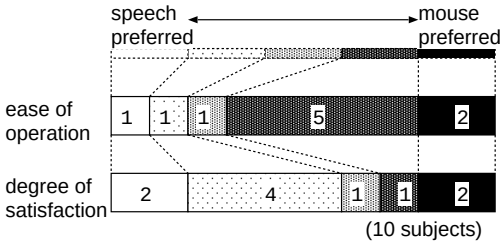

Figure 4. Speech input vs. Menu selection by mouse

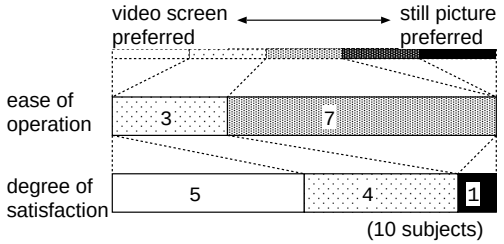

Figure 5. Video image on large screen vs. Still pictures on computer display

For now, we contend that providing satisfaction to users through greater interaction is highly important, and our system realizes natural and robust interaction while users keep the initiative.

REFERENCES

[1] A.Waibel, B.Suhm, M.T.Vo, and J.Yang. Multimodal interfaces for multimedia information agents. In *Proc. IEEE-ICASSP*, pages 167–170, 1997.

[2] S.Furui. Prospects for spoken dialogue systems in a multimedia environment. In *Proc. ESCA workshop on Spoken Dialogue Systems*, pages 9–16, 1995.

[3] A. Pentland. Smart rooms, desks, and clothes. In *Proc. IEEE-ICASSP*, pages 171–174, 1997.

[4] R.A.Bolt. Put-that-there: Voice and gesture at the graphics interface. *ACM Computer Graphics*, 14(3):262–270, 1980.

[5] T.Kawahara, M.Araki, and S.Doshita. Comparison of parsing and spotting approaches for spoken dialogue understanding. In *Proc. ESCA workshop on Spoken Dialogue Systems*, pages 21–24, 1995.

[6] T.Kawahara, T.Kobayashi, K.Takeda, N.Minematsu, K.Itou, M.Yamamoto, T.Utsuro, and K.Shikano. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *Proc. Int'l Conf. on Spoken Language Processing*, pages 3257–3260, 1998.

[7] J.Flanagan and I.Marsic. Issues in measuring the benefits of multimodal interfaces. In *Proc. IEEE-ICASSP*, pages 163–166, 1997.